NAEP Newsletter #7

SCORING WRITING

A recent observation of a 2004 MEA anchor selection meeting followed by the initial scorer training for an 8th grade prompt led to some reflection of what I have seen over 10 years of involvement with NAEP scoring sessions for constructed response items. This issue of the *Newsletter* is an accounting of the intersections of MEA and NAEP writing scoring practices.

-J. H. Kennedy, NAEP State Coordinator for Maine

NAEP Writing

NAEP assessments are built to specifications established in subject frameworks, which are developed by committees of teachers under the supervision of assessment specialists and approved by the National Assessment Governing Board (NAGB). The current NAEP Writing framework was first implemented in 1998. NAEP frameworks generally have a life span of 10 years.

The previous NAEP Writing framework used a more analytical scoring guide than it does now; in that system of scoring, readers looked for the presence or absence of specific features of the response. In the current approach, scorers develop an overall impression of the ability of the writer. In the pre-1998 assessments, the whole set of student responses for a given prompt was scored twice (once for topic development and organization and a second time for mechanics) using different scoring guides. Overall, the scoring of pre-1998 NAEP writing was more aligned methodologically with the scoring of written responses in other NAEP assessments (e.g., Reading, History, Science) than it is now.

The current NAEP writing assessment uses a six-point modified holistic scale. Both topic development and language use are considered in assigning a score based upon the overall impression of the scorer, which is based upon the training that scorer has received. A student completes two prompts for a possible score of 12, based upon a single reading of each prompt. A portion of all of the responses is read a second time to assess reader reliability, but scores for second readings are not combined with scores for first readings.

When an assessment framework changes as much NAEP Writing's did in 1998, the *Trend Line* is said to be broken, and NAEP will no longer report data for the two assessments together. This is to prevent the invalid comparison of two sets of scores having very different meanings.

Similarly, MEA and NAEP scores cannot be compared directly, either in terms of achievement levels or in the scores themselves. What can be compared are trends and gaps in performance.

MEA Writing

NOTE: The current MEA writing assessment contains one traditional prompt and two new tasks per grade. The scoring of the prompt is most similar to the scoring of NAEP writing. All references to MEA writing in this document are to the traditional prompt.

In the 2004 MEA scoring of writing, two scorers assign two scores (using a scoring guide with two different components) to each student response, for a possible total of 20 points. The first score of one to six points is for the stylistic and rhetorical aspects of writing and for topic development; the second score of one to four points is for standard English conventions.

Similarities in NAEP and MEA Scoring of Writing

Both assessments use a modified holistic approach for scoring student responses; the readers balance their evaluation of designated features of the writing in order to develop an overall impression of the ability of each student to write.

With both assessments, scorers are trained to evaluate students responses as first draft writing in which one would not expect to find the perfection of a polished finished product. Scorers are trained to view the prompt as a 'springboard' for writing; this means they look for a link (even a tenuous one) between students' responses and the prompt. Thus students are able to respond using their own background knowledge and experience. Specified outside knowledge is not required to complete this assessment successfully; the ability to write clearly about one's thinking is.

Both assessments may include prompts in a variety of modes; however, these types of writing are not necessarily considered exclusive of each other by either of the assessments. NAEP topics may be narrative, persuasive, or informative; MEA topics may focus on narration, exposition, description, or persuasion (except for persuasion in 4th grade). It is pointed out for scorers of both assessments that an exposition or informative piece, for instance, may include features (or sections) of narration and/or persuasion.

For both assessments, the scorers are trained first on a scoring guide and then on *Anchor Papers*, which illustrate expectations of student performance at each level of the scoring guide. Scorers refer to both of these resources while scoring each student response, with the anchor papers serving as the primary source of guidance.

Scoring of both assessments is done by the staff of professional scoring centers established by national testing companies; the MEA, beginning with the 2003-04 school year. These staff may be educators, past or present, or they may be from other professions. All are selected for their ability to apply reliably the scoring system on which they are trained to the student responses they score.

Both assessments train scorers on paper samples and then have them score actual student responses on computer networks. Both assessments use a combination of current year and past student responses in training scorers. For MEA, the benchmark year for scoring writing is 1998-99. Previously scored responses, sometimes on the same topic and

sometimes on other topics, have been used over the intervening years to link current scoring practice to the benchmark year.

NAEP typically has used responses on the same topic produced and scored in previous years in a variety of ways to check current scoring practices of scorers; the responses might be included in the training materials, or they might be inserted into the live scoring as a means of assessing the similarity of scorer decisions in the current year and previous ones. In this way and others, scorers for both assessments are trained specifically to prevent *drift* and *bias*.

Scorer drift occurs when a group of scorers agree with each other but their agreement is moving away from decisions made by other groups of scorers in previous years or by their own decisions in previous days. For NAEP, it is prevented by performing statistical analyses (*t-tests*) on the overall pattern of the scoring team decisions; if these tests indicate that scorers in the current year are not behaving statistically in the same way as scorers of the same material in previous years, the scoring may be stopped and the scorers retrained.

Bias can be caused by scorer reaction to factors as variable as the point of view of the student writing the response or the handwriting of the student or length of the response. This is prevented in part by careful selection of anchor papers by the scoring center staff trainers and a committee of Maine teachers in the case of the MEA (or, for NAEP, a combination of trainers and content specialists). Summaries of the trends of current individual scorers (scoring high or scoring low consistently, for instance) are periodically reviewed by scoring supervisors.

Scorers for both assessments are supervised in small groups by a *Table Leader*, who has received additional training in scoring the assessment. The scorers and table leaders are supervised by a *Chief Reader* for the MEA. In NAEP scoring, each table leader works collaboratively with a different scorer trainer. The NAEP trainers report to a *Writing Content Specialist*. Persons working at each of these higher levels of supervision are chosen for their extensive experience in scoring many different types of writing assessments.

Table leaders and the chief reader/trainers monitor scorers by observing their work in real time (*Back Reading)* and cumulatively with on-screen reliability tools that indicate patterns of agreement with other scorers. For NAEP, the writing content area specialists monitor reliability of scoring groups but generally do not get involved with individual scorers. MEA chief readers closely monitor all scorers, either directly or indirectly (through reports by table leaders).

When scorers in either assessment are unsure of what score to give a response, they are trained to look first at the anchor papers and then to consult with their table leaders, who in turn may consult with the chief reader/trainers. For NAEP, trainers may, in turn, confer with their writing content specialist. NAEP's holistic scoring procedures have been influenced to an extent by Educational Testing Service (ETS) writing specialists

who developed Praxis (formerly NTE), GMAT, and GRE writing measures and their scoring procedures whose development started in the late 1960's. The MEA writing assessment's development has been somewhat influenced by NAEP's writing assessment.

NAEP's pre-1998 writing assessments used a hierarchy similar to the current MEA system; i.e., a chief reader supervised table leaders, and the entire group scored one prompt at a time. Since then, the growing size of the population assessed by NAEP has made it necessary to find more efficient ways of processing increasingly larger volumes of student work.

Logistical Considerations

Scorers usually see only handwritten responses to the NAEP assessment, but MEA scorers began this year to see some typed responses at the 8th grade level for writing as the *MEA Online* assessments are gradually being phased in. Handwritten responses for both assessments are scanned into e-files for display onscreen to scorers. NAEP scorers receive responses online in packets; 25% of these packets are then circulated to a second scorer on the same small team scoring a particular prompt.

MEA scorers receive an initial set of 10 of responses that are examined carefully by scoring supervisors for quality control and then single responses that are circulated randomly to anyone scoring writing. MEA writing scorers work as a large group on one prompt at a time, so the pool of available second readers is quite large compared to NAEP. NAEP scorers work in small groups (*tables* or *teams*), each working on a different prompt at the same time, so the available pool of second readers is smaller than for MEA. The second reading of an entire packet of NAEP responses is done by one second reader, further reducing the random selection of a different second reader for each response. NAEP 4th, 8th and 11th grade assessments are scored simultaneously in small groups of scorers in this way; in MEA, the responses for each grade are scored separately by the whole group.

Differences in NAEP and MEA Scoring of Writing

A lot of the differences in scoring of the two assessments will naturally have to do with the differences in the student responses with which scorers are presented, and that is a natural consequence of differences in the stimuli to which students are responding.

While some NAEP and MEA and prompts may appear to be quite similar, they are generated by different frameworks based upon different kinds of standards. MEA standards are performance standards taken from Maine's *Learning Results.* There are no federal national standards in writing (or any other subject) because curriculum standards are controlled by each state per federal law. MEA scores indicate the degree to which students meet the state's achievement expectations. NAEP scores have meaning only in context, one of which is the *Achievement Levels*, performance standards established by review of student responses to the assessment and the assigning of their score point levels

as being representative of levels of *proficiency*. Unfortunately these standards have yet to be validated and remain provisional.

Average scaled scores are the actual result of NAEP assessments. As we have said, they must be put in context to have meaning. This can be done through subgroup analysis. For example, NAEP scores are useful for comparing the performance of different populations of students (males and females, for example) or for tracking the improvement of performance of single groups of students over time (impoverished male 8th grade students in rural areas of Maine, for example). They have meaning only in these contexts and do not provide benchmarks against any universal writing standard.

It is important to remember that NAEP scores represent predictions of a sophisticated statistical model of the expected performance of large groups of students based upon the actual performance of smaller groups of the same types of students. Also, NAEP frameworks represent a cross-section of the varieties of writing instruction found in this country's schools, not a comprehensive survey of all writing instruction in the country. For example, in some parts of the country, persuasive writing may not be taught prior to 8th grade, but the students administered the NAEP writing assessment for that grade may be presented with persuasive writing prompts.

On the other hand, MEA prompts can solicit only performance that is described in the *Learning Results* for the grade assessed, and it is assumed that instruction throughout Maine covers this content.

With this in mind, an examination of the subgroup differences in scores on MEA and NAEP of boys and girls, for instance, reveals additional ways of looking at the differences between the two assessments.

The substantial gender gap in reading and writing is well-known and universal, according to national and international assessments. This is not an example of scorer bias; in fact, there is much debate over what causes the gender gap, but it is most likely something inherent in either the students or the assessments or a combination of the two. Using a statistical concept known as *effect size*, it is possible to compare gaps in different assessments using different scales for creating scores. PIRLS is the international assessment of 4th grade Reading. The chart below compares the size of the gap in PIRLS, NAEP/Maine, and MEA scores on the most recent results available for each (2000-2003).

Note: Effect size is calculated by dividing the size of the gap between subgroups by the standard deviation (range of scores) of the entire population assessed.

	PIRLS	ME NAEP	MEA
4th Math	(no data available)	0.11 (males)	0.00 (males)
8th Math	(no data available)	0.06 (males)	-0.11 (males)
11th Math	(no data available)	(no data available)	0.12 (males)
4th Reading	0.21 (females)	0.14 (females)	0.37 (females)
8th Reading	(no data available)	0.37 (females)	0.58 (females)

11th Reading	(no data available)	(no data available)	0.86 (females)
4th Writing	(no data available)	0.61 (females)	1.41 (females)
8th Writing	(no data available)	0.42 (females)	1.09 (females)
11th Writing	(no data available)	(no data available)	1.24 (females)

Note: Scoring for the 2004 MEA had not been completed at the time of this writing.

We can see from the analysis above that the gap between females and males in reading and writing is always in favor of the females; in addition, the gap is much larger for writing than for reading. Further, the gap is much larger on the MEA than on the NAEP at 4th and 8th grade. Whatever the reason for the gap, these data should be seen as an indication that the MEA is able to delineate the problem more clearly and not as an indication that the MEA is harder than the NAEP. A look at the differences in the conditions under which students work on these assessments helps to clarify this point.

While scorers of both assessments are instructed to view all student responses as first draft writing on demand without prior knowledge of the topic, the expectations of student performance on each assessment differs somewhat because of the conditions under which each is administered. Students taking the MEA can have up to 70 minutes (45 minutes with an optional 25 more) to complete their work on a prompt; in addition, these students may use reference sources available in the classroom. Students taking the NAEP have 25 minutes to complete the writing task and may not use any classroom resources.

The level of expectation of the scorer in holistic scoring is set by the samples of work used in training to calibrate the performance expected to achieve each score point level. The level of work produced under the two different conditions of the MEA and the NAEP lead naturally to different expectations. All other things being equal, the MEA work should naturally be of higher quality for a given student because of the conditions under which it is produced.

Another interesting distinction along the same lines, between MEA and NAEP training, is that MEA anchor papers are presented to scorers from the lowest to the highest rated, while NAEP anchor papers are presented from the highest to the lowest rated. NAEP scorers are instructed to 'score supportively' and to 'give the highest score the scoring guide allows.' This means that the scorer should look for what the response does well rather than what it is lacking.

A comparison of key words in MEA and NAEP scoring guides reinforced in training observed suggested at first that the overall holistic impression of a MEA scorer is somewhat consistently a point lower than a NAEP scorer.

As a NAEP trainer, I used the following catch phrases to describe levels of performance:

Upper-Half Response
A **6** shows consistent variety and complexity.
A **5** shows emerging variety and complexity.

A **4** is clear and simple.

Lower-Half Response
A **3** is fatally flawed (in one of several possible ways)
A **2** is fragmentary or fades in and out of coherence.
A **1** is minimal or incoherent.

For MEA scorers, a "barebones response" would receive a 3, while NAEP scorers would give such a clear and simple response a 4. MEA scorers are encouraged to make an initial decision as to whether a given response is closer to a 2 or a 5. For NAEP scorers, the "high-low" decision would be centered around scores of 3 and 4. Nascent characteristics are given a relatively high value in NAEP scoring, while MEA scoring appears to look for a more developed sense of the writer's abilities in a response. Also, the NAEP scoring guide rewards variety in sentence structure with a 5, while the stylistic/rhetorical part of the MEA guide indicates that this is a characteristic of a 4. Thus a 5 response for a NAEP scorer might be a 4 for a MEA scorer.

However, making such a judgment about MEA and NAEP scoring is complicated by differences in the content of the scoring guides. The MEA writing assessments place the concept of *voice* in a pivotal role in determining the development/style portion of the score. In MEA scoring, voice and tone is seen to emerge through language and style choices which are the final consideration in a hierarchy described by the 2004 chief reader's training charts reproduced below.

Chart 1 (*read down*)

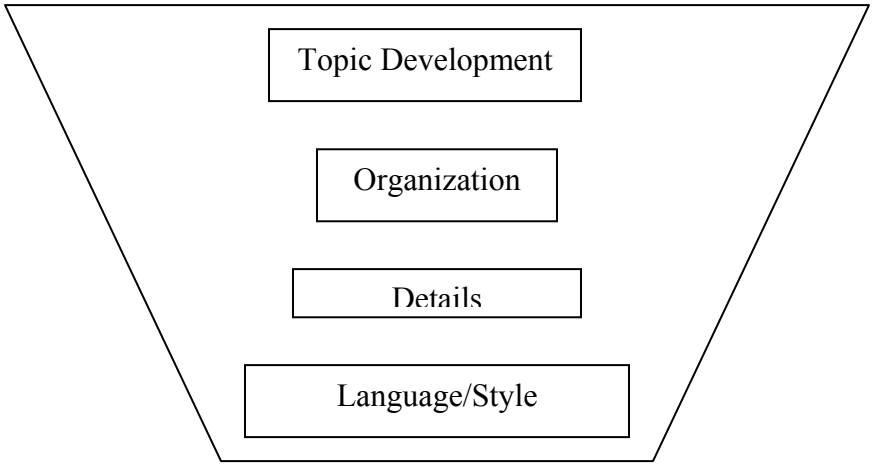
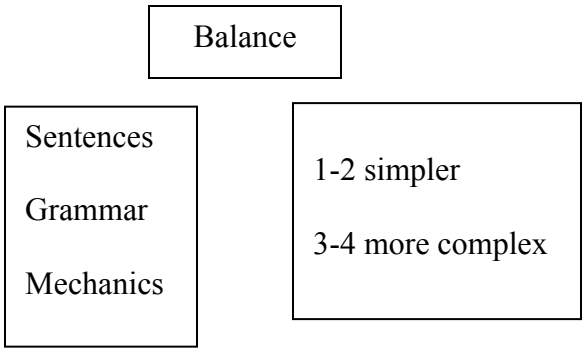


Chart 2:



Using these charts, the chief reader described the holistic scoring process for the one to six point style/development judgment (Chart 1) to be a hierarchical one, rising from considerations of language and style through weighing the accumulation of detail to assessing the organization of the material to an overall sense of the quality of the topic development to arrive at a 'balance' of elements that produced an equitable score.

Then, on the second scoring (for mechanics, Chart 2), a balance of impressions of the quality of the sentence structure, or grammatical error or lack thereof, and of the adherence to standard conventions in mechanics leads to a bi-level judgment of simplicity (1/2 points) or complexity (3/ 4 points) in language use and then a refinement to the single score.

NAEP scoring judgments largely avoid this type of content issue in favor of sentence structure and organization of material. For a NAEP scorer, the complexity of a response would be the natural consequence of having a great deal of relevance to say and saying it effectively. The resulting communication could be devoid of voice and still score at the highest levels of the rubric. The result of these differences is that some MEA anchor papers would receive a higher score from a NAEP scorer and some would receive a lower score because the judgments of both scorers are being made holistically, but with a different mix of criteria being balanced out to come to a scoring decision.

Similarly, MEA scoring also looks more specifically at types of error in language use for its second category scoring (mechanics), while NAEP scoring looks at language use as a function of communication and assesses its impact largely in terms of interference with communication along a continuum from none to too much and holistically in combination with features of development. Thus, NAEP scoring is done by taking a broader look at specific features of student performance than does MEA scoring.

In NAEP scoring, topic development can be seen as a series of structures built one upon another, starting at the sentence level and progressing through transitional devices to

arrive at an overall coherence. MEA scoring is based upon the hierarchical series of judgments described in the charts above, which start with topic development and end with style, organization and details being intermediary steps.

While a principle of holistic scoring, used by both assessments, is to permit the relative strength in one aspect of the performance assessed to offset the relative weakness in another aspect of that performance, the structures of the MEA and the NAEP *rubrics* (systems of scoring) result, in particular, in different ways in which the ability to control language is assessed by each.

The NAEP's method of assigning one holistic score for topic development and language use permits relative weakness in the one to offset relative strength in the other (and vice versa). The MEA's method of assigning separate scores for development and language does not permit this leeway. Neither approach seems more appropriate when considered in its context; the MEA is measuring achievement of specific curriculum standards, while the NAEP is taking a broad measure of the ability of students to communicate in writing. The MEA's results are directed towards the improvement of instruction; the NAEP's results are directed towards assessment of progress and the resulting modification of educational policy at a state or national level as a result. Each has its place a comprehensive assessment system that respects both state and national priorities.

The author was Chief Reader for the scoring of the 1996 NAEP Writing assessment and a scorer trainer for the 1998 and 2002 writing assessments. In addition, he served for two years on the ETS/NAEP team that developed the current NAEP writing assessment. He was a NAEP scorer trainer in other subjects from 1993 to 2002.

Samples of the MEA and NAEP scoring guides may be obtained from J. H. Kennedy, NAEP State Coordinator for Maine, 23 State House Station, Augusta, Maine 04333. 207-624-6636 (john.kennedy@maine.gov)

NAEP 2005 assessments in Reading, Mathematics, and Science will be administered in January and early February at 4th, 8th, and 12th grade. State results will be reported for 4th and 8th grade. About 60% of Maine schools will be asked to participate, and those selected will be notified this summer. The No Child Left Behind Act requires state participation. The last results for Reading and Mathematics were 2003; Science, 2000.